Navy Personnel Research and Development Center

San Diego, CA 92152-6800 TN 88-13 January 1988

Test of a Probabilistic Sampling Critical Task Selection Model for Performance Testing

DEPARTMENT OF THE NAVY
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152-6800

3900
Ser 62/54
14 JAN 1988

From: Commanding Officer, Navy Personnel Research and Development Center

Subj: **TEST OF A PROBABILISTIC SAMPLING CRITICAL TASK SELECTION MODEL FOR PERFORMANCE TESTING**

Encl: (1) NPRDC TN 88-13

1. The Navy Job Performance Measurement Program is an outcome of the Navy Performance-Based Personnel Classification Subproject (Z17701.001). Both efforts constitute significant contributions to the Joint-Service Job Performance Measurement/Enlistment Standards Project. The Joint-Service Project has been mandated by Congress to link enlistment standards to job performance, which can be considered a landmark research thrust of the armed services. The present research has been funded primarily under Project Element Number 63707N (Manpower Control System Development) and Project Number Z1770 (Manpower and Personnel Development).

2. This report details the testing of a probabilistic sampling model for assessing the performance of first-term Navy Radiomen (RM). The information it contains is intended to benefit the research and operational RM communities. Ultimately, the outcome of the project will benefit the armed services, military and civilian research communities, and applied industrial organizational psychology in general.

JOHN J. PASS
By direction

Distribution:
Office Assistant Secretary of Defense (OASD) (FM&P)
Chief of Naval Operations (OP-135L)
Defense Technical Information Center (DTIC) (2)

NPRDC TN 88-13 January 1988

Test of a Probabilistic Sampling Critical Task Selection Model for Performance Testing

Steven E. Lammlein, Ph.D.
Norman G. Peterson, Ph.D.
Rodney L. Rosse, Ph.D.
Personnel Decisions Research Institute

Reviewed by
Herbert George Baker, Ph.D.

Released by
John J. Pass, Ph.D.
Director, Personnel Systems Department

Navy Personnel Research and Development Center
San Diego, California 92152-6800

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NPRDC TN 88-13	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Personnel Decisions Research Institute 6b OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Navy Personnel Research and Development Center
6c ADDRESS (City, State, and ZIP Code) Minneapolis, MN 55414	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 63707N; PROJECT NO: Z1770; TASK NO: Z1770.001; WORK UNIT ACCESSION NO:

11 TITLE (Include Security Classification)
Test of Probabilistic Sampling Critical Task Selection Model for Performance Testing

12 PERSONAL AUTHOR(S)
Lammlein, S. E., Peterson, N. G., Rosse, R. L.

13a. TYPE OF REPORT	13b TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Interim	FROM 1982 TO 1986	1988 January	26

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Critical tasks, task selection, probabilistic sampling model, job performance measurement

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This study tested a probabilistic sampling model for assessing the performance of Navy radiomen (RM). Five steps were involved: (1) a Monte Carlo investigation tested the effects of simulated alternate task weighting systems; (2) RM job experts determined the actual task weighting systems; (3) task samples were drawn; (4) experts screened the task samples; and (5) the final task sample was selected. Results indicated that the probabilistic sampling model was considerably influenced by the random luck of the draw. The results were not positively viewed by RM experts. Suggestions for probabilistic sampling procedural modifications are presented.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Baker, Herbert George	22b. TELEPHONE (Include Area Code) (619) 553-7639
	22c. OFFICE SYMBOL Code 62

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete.

SUMMARY

Problem

Many job performance measures--with "hands-on" or work sample measures being the clearest example--are based on job tasks. Such measures are developed by selecting a subset of tasks from the job as a whole to form the sample. One can select different sets of such "critical" tasks that will vary in the degree to which they capture the essence of the job as a whole. The rationale and procedure by which critical tasks are selected is thus an important issue.

Purpose

The purpose of this study was to test a probabilistic sampling model for assessing the performance of Navy Radiomen (RM).

Approach

There were five primary research steps in this study. (1) A Monte Carlo investigation tested the effect of simulated alternate task weighting systems, (2) RM job experts determined the actual task weighting systems, (3) task samples were drawn, (4) experts screened the task samples and (5) the final task sample was selected.

Results

RM experts found the task samples unacceptable for performance testing because they did not accurately reflect the first-term RM job. Several suggestions for probabilistic sampling procedural modifications are presented.

Conclusions

1. The primary assumption of the model--that all tasks performed by those holding the job title represent a coherent domain from which it is useful to conceive of a domain score--is flawed given the diverse nature of all but the most simple jobs. It is quite difficult to draw a valid analogy between a population of experimental subjects and a population of tasks for most jobs.

2. Given that task samples must generally be small due to administrative constraints, these samples may be subject to considerable error in representing the entire "domain" of tasks if drawn randomly.

Recommendations

It is suggested that future research in the area of probabilistic sampling address a number of methodological alternatives proposed in this report.

ACKNOWLEDGEMENTS

The authors wish to thank a number of persons for their help in carrying out this research study. From the Human Resources Research Organization, John Ziemak conducted workshops and Jim Harris provided useful advice at several points in the project, including helpful comments on an earlier draft of this report. Marvin D. Dunnette of Personnel Decisions Research Institute also provided useful comments on the earlier draft. Pamela Kidder of the Navy Personnel Research and Development Center was a most valuable help in identifying job experts and setting up workshops. Finally a great debt of thanks is owed to the Navy radioman job experts--Daniel Dick, Henry Kelley, Robert Nicholson, W. D. Nickson, Robert Van Auken, and Glenn Vierra--who very competently and patiently participated in formulating the weighting systems and the review of the task samples.

CONTENTS

LIST OF TABLES

INTRODUCTION

Measures used to assess job performance are by necessity samples, as we cannot carefully observe and objectively evaluate all the performance of any incumbent (Lammlein 1986). A most important issue in job performance testing is that of job sampling: On what basis are certain parts of the job to be selected for performance testing over others? The way in which jobs are sampled has significant implications for the interpretations that may be legitimately attached to job performance scores.

Many job performance measures--with "hands-on" or work sample measures being the clearest example--are based on job tasks. Such measures are developed by selecting a subset of tasks from the job as a whole to form the sample. One can select different sets of such "critical" tasks that will vary in the degree to which they capture the essence of the job as a whole. The rationale and procedure by which critical tasks are selected is thus an important issue.

Background and Conceptual Overview

The Navy has been developing three types of job performance measures for the first-term radioman (FTRM) job: hands-on and written job knowledge tests and rating scales. These measures are based on a set of 22 critical tasks identified for the FTRM job. To identify these critical tasks, an extensive job analysis survey with 124 tasks[1] was administered to both FTRM and their supervisors to learn which tasks are more important, more frequently performed, etc. For more information on this part of the study, the reader may consult Lammlein and Baker (1987). The same collection of 124 job tasks was used in the present study.

The rationale for selecting critical tasks on the basis of considerations such as importance and frequency can be traced to test theory and utility formulations. It underlies many performance measurement studies, which often use either informal panel discussions or more structured job analysis surveys to identify the more salient tasks or other components of the job. It is referred to in this report as the "utility" model for task selection.

An alternate model, based more on statistical sampling theory, has been advocated for pilot testing by the Committee on the Performance of Military Personnel, Commission on the Behavioral and Social Sciences and Education of the National Research Council (1986). For purposes of this report it will be termed the "probabilistic sampling" model. A brief overview of the two models follows.

The Utility Model

The essence of the utility model is the selection of critical tasks that are most salient to the job. In employing the utility model, one must first develop some operational definition of "salience." This definition will necessarily depend on the purpose of collecting the performance data. Most commonly, judgments of task performance, frequency, or some combination of the two are used. This information is then collected via methods such as job expert panel discussions or task surveys.

---

[1] This task list is available from NAVPERSRANDCEN."

The rationale for this utility model is straightforward: The more salient a task is to a job (i.e., the more important it is or the more frequently it is performed) the stronger the consequences associated with performance/non-performance of that task for the organization. Therefore, the more utility there is for the organization in assessing performance on it--thus, the designation as the "utility" model.

In the FTRM effort, several types of job analysis information were used to operationally define "salience" and select the critical tasks. The three most prominent features were task importance (including both importance for mission success and frequency of performance), task variability (the percentage of time the task is performed correctly and how complicated it is to perform), and task predominance (the percentage of FTRM supervisors who supervise the task and the percentage of FTRM who perform it).

All this information was collected from the job analysis survey of FTRM supervisors and incumbents. The goal in task selection was to select those tasks receiving high ratings on the importance and frequency scales and moderate ratings on the percentage performing correctly and "complicated to perform" scales. In addition, the selected tasks were those performed/supervised by a relatively high percentage of incumbents/supervisors.

These data-based selections were subsequently reviewed and slightly revised by a committee of Navy job experts who considered factors such as feasibility of testing, future needs of the Navy, etc. These factors are most difficult to assess in a job analysis survey.

This is an example of the utility model, supporting the rationale for using the task importance and task predominance information. In addition, task variability was included in the operational definition because the purpose of the performance testing was to compare different methods of assessing job performance. Obviously, it is useful to make such a comparison on tasks for which there is a performance variance.

The Probabilistic Sampling Model

A probabilistic sampling is based on a somewhat different rationale: The score that one receives based on evaluations of performance on the critical tasks should be an unbiased estimate of the score that one would receive if tested on all job tasks, allowing for some weighting of the task scores based on salience of the tasks. This approach thus primarily emphasizes representativeness of the task scores in contrast to the previously described approach, which gives primary emphasis to the utility consequences of task selection.

At the heart of the probabilistic sampling model is a probability distribution for job task selection and the random sampling of critical task samples using that distribution. The probability distribution reflects relevant information about each task such as task salience (e.g., importance, frequency), task category membership, etc. The actual probability values are a function of a weighting system, dictated by job experts, which combines this relevant task information in a way that is meaningful to determining a theoretical performance score or scores.

For example, job experts may believe that the most useful performance score would be that obtained if a FTRM could be tested on all 124 job tasks. This summary score would be computed by weighting each task score by a value consisting of the importance mean plus the frequency mean for that task (from the job analysis survey), summed across

all tasks (it will be assumed for now that the task performance variances are equal and the covariances between tasks are equal as well). The probability of task selection for a sample from that 124 should thus be based on a sum of the importance and frequency means, and this additive model would then be used to derive a probability distribution for task sampling.

Job experts decide on the composition of a score or scores of interest, and their judgments are reflected in the formation of a probability distribution for the sampling of tasks. It is believed by proponents of this model that random sampling of tasks from such a distribution will produce performance scores representative of the hypothetical value that would be obtained if performance could be tested on all tasks in the performance domain of the job.

Once the probability distribution is formulated, 150 random samples of tasks are drawn of the desired size for the critical task sample (the number of tasks drawn in a given situation will vary depending on administrative constraints and other factors influencing how many tasks can be feasiblv tested). Job experts then review the 150 samples and delete at most 10 percent of them for reasons of unacceptability. Of the remaining samples of tasks, one is chosen at random as the "final" critical task sample for performance measurement purposes.

Choice of Model Depends on Purpose

Note that both models use similar task information, but in different ways. The two models are based on quite different assumptions about what one wishes to generalize to from performance on a sample of tasks. The probabilistic sampling model, derived from statistical sampling theory, is predicated on the assumption that one will wish to generalize to performance on the entire domain of tasks. The utility model attempts to generalize only to performance on the more salient aspects of the job, with salience being related to worth of task performance to the organization.

Purpose of This Study

The research described here was essentially a test of the probabilistic sampling model as an alternative to the utility model that had previously been employed in the Navy RM study. The procedures of the probabilistic sampling model were followed as closely as possible, with one exception. Rather than using one team of job experts to determine the weighting system and delete task samples, two were used so that results could be compared. In addition, a preliminary step--a Monte Carlo investigation--was added to the study to assess how the composition of task samples was influenced by different weighting systems.

METHOD, RESULTS, AND DISCUSSION

There were five primary research steps in this study: (1) the Monte Carlo investigation to test the effect of simulated alternate task weighting systems, (2) the determination by job experts of the actual task weighting systems, (3) the drawing of task samples, (4) expert screening of the task samples, and (5) selection of the final task sample.

Monte Carlo Simulation

Job tasks differ with respect to importance, frequency of performance, etc. The probabilistic sampling model allows for these differences to be reflected in the probabilities of task selection through what is essentially a weighting system job experts assign to the task information. An important question is suggested by this: Just how important are different weighting systems? If the weighting system is widely varied, how will the composition of the task samples change?

Method

Probability distributions based on five different weighting systems were constructed: (1) uniform (no different weighting), (2) importance weighting only, (3) frequency weighting only, (4) weighting by a simple sum of importance and frequency, and (5) weighting by the product of importance and frequency. This was accomplished by first applying the weights of each system to the mean importance and frequency values for the task items as obtained from the job analysis survey in Lammlein and Baker (1987). To form the probability values, these results were then rescaled with the constraint that the probabilities for all 124 tasks must sum to 1.0 for each weighting system.

For each probability distribution, 150 random samples of 15 tasks each were drawn using a computerized random sampling procedure. Sampling of each set of tasks was done with replacement. Fifteen was used as the task sample size because this is the maximum number of tasks that can typically be tested in a hands-on format.

A useful dependent variable for this investigation is the inter-sample task overlap; simply, the percentage of tasks common to a pair of task samples. A baseline overlap value can be established using the task samples based on the uniform weighting, in which each task has an equal probability of selection. The overlap between different samples chosen by a non-uniform weighting system should be higher than this baseline if the weighting does in fact significantly impact task selection. It is intuitively plausible that the more marked the differential weighting of selection probability (i.e., the more marked the difference in probability of selection between the task least likely to be selected and the task most likely to be selected), the more overlap there should be among task samples. Thus, the more extreme weighting systems (e.g., the multiplicative weighting of importance and frequency) should produce higher inter-sample overlap values.

Results

Table 1 presents summary statistics for task probability distributions for the five weighting systems. The mean probability value of each distribution is the same--1.00/124 = .00806. The other information in Table 1, however, is useful in that it indicates that the different weighting systems did influence the shapes of the probability distributions for task selection. One sees, for example, that the ratio of largest/smallest probability values for the multiplicative system (importance X frequency) is almost four times that for the uniform weighting. This is reflected as well in the standard deviations of the weights.

Overlap analyses were carried out by computing the proportion of overlap between each task sample drawn by a given weighting and (1) the remaining 149 task samples drawn using the same weighting, and (2) the 150 task samples drawn by each of the four other weightings. In addition, the mean importance and frequency values were computed for each task sample. Table 2 summarizes the overlap results. The values in this table

Table 1

Summary of Probability Distributions Used in Monte Carlo Investigation of Effect of Different Weighting Systems

	Uniform	Importance	Frequency	Importance + Frequency	Importance x Frequency
Mean	.00806	.00806	.00806	.00806	.00806
SD	.00000	.00132	.00209	.00116	.00241
Maximum	.00806	.01099	.01216	.01049	.013 6
Minimum	.00806	.00507	.00417	.00540	.00353
Ratio[a]	1.00	2.17	2.92	1.94	3.90

[a]Ratio of maximum to minimum probability weight.

Table 2

Summary of Overlap Analyses for Monte Carlo Investigation of Effect of Different Weighting Systems

	Uniform		Importance Only		Frequency Only		Importance + Frequency		Importance x Frequency	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Uniform	.120	.079								
Importance	.120	.079	.125	.080						
Frequency	.122	.080	.122	.079	.130	.081				
Imp. + Freq.	.122	.080	.123	.079	.125	.080	.125	.080		
Imp. x Freq.	.121	.079	.125	.081	.130	.082	.126	.081	.132	.081

Note. Values in this table are based on proportions of overlap between task sets.

are the means and standard deviations of overlap proportions for task samples chosen by weighting combinations. As an example, consider the mean of .120 for the combination of uniform weighting with itself. This value means that, on the average, 1.8 tasks (12% of the 15 tasks in a sample) were shared by combinations of two task samples each drawn from the uniform weighting distribution. Of primary interest here are the diagonal values in this table, which should, if weighting does in fact make a difference, increase as one moves from the uniform to the more differential weightings. It is apparent from Table 2 that while such an increase is present it is of little practical significance. The largest

mean proportion is .132, or 1.98 tasks in common for the multiplicative combination of importance and frequency. This represents only a miniscule increase over the 1.8 tasks in common observed for the uniform weighting.

A similar result is evident from Table 3. This table depicts the composition if the task samples drawn from each weighting in terms of importance and frequency means (from the FTRM job analysis). For example, the value 3.40 for importance with the uniform weighting means that the average sample of 15 tasks drawn with that weighting had a mean importance rating across tasks of 3.40. One should see the importance and frequency means increasing to the extent that they are differentially emphasized in the weighting systems. In fact, the highest mean for importance is observed for the importance weighting, and likewise for frequency. However, again, the practical significance is not marked.

Table 3

Summary Statistics for Importance and Frequency for Task Sets Selected in Monte Carlo Investigation of Effect of Different Weighting Systems

	Importance		Frequency	
	Mean	SD	Mean	SD
Uniform	3.40	.132	2.9	.193
Importance only	3.48	.142	2.95	.184
Frequency only	3.41	.147	3.11	.209
Importance + Frequency	3.43	.135	3.02	.176
Importance x Frequency	3.45	.146	3.09	.209

Discussion

In the probabilistic sampling model, two factors determine which tasks will be chosen for performance testing. One is task information such as importance, frequency of performance, etc. as used in some weighting system. The other is the luck of the draw as exercised through random drawing of tasks. The question addressed by this Monte Carlo study was: What is the relative influence of each of these two factors on outcomes of the probabilistic sampling model?

The overall conclusion is that the practical effect of the differential weighting systems on task sample composition was rather small. Put another way, task samples did not become substantially more alike--in terms of shared tasks--as the probability weighting was made to be more differentiated across tasks.

In the probabilistic sampling model, 150 tasks samples are narrowed to 135 on the basis of expert judgment. Then, each of those 135 has an equal probability of selection as the "final" sample for testing. For the parameters considered in this study, (124 tasks in total, samples of 15 tasks, etc.), the luck of the draw would play a most significant role in the composition of the final task sample. Using even the most extreme weighting

considered, the average number of shared tasks between samples is only about two. Thus, on the average, 13 out of 15 tasks differ between samples that have an equal probability of selection. Whether such a result is tolerable depends on the generalizations one wishes to make from the task sample as well as the characteristics of the job.

It is useful to emphasize a limiting point on the conclusion of little weighting effect. The effect of weighting could reasonably be expected to vary with factors such as the number of tasks to be sampled; the overall number of tasks from which to draw; and the differentiation between tasks in terms of importance, frequency, and other job information.

Task Weighting by Job Experts

The probabilistic sampling model calls for job experts to formulate how task information is to be used in the formation of a probability distribution for task sampling. The set for this rather abstract step is presented in terms of the formation of an overall "score" that would be of interest to the job experts if radiomen could be tested on every job task. Job experts indicate how these individual task performance scores should be combined with other task information to yield the overall score(s) of interest to them. Thus, some weighting of task information is implied.

Method

In this study, two teams of job experts independently formulated weighting systems, with one team meeting in San Diego, California, and the other in Alexandria, Virginia. As Table 4 shows, the participants in these two workshops (1) are all very experienced with the radioman job, and (2) represent a number of different perspectives in terms of current duty assignments.

Table 4

Radioman Job Experts Participating in Formulation of Weighting Systems

	Grade	Current Assignment	No. of Years as Radioman or Radioman Supervisor... On Ship	On Shore	Total
		Team 1. San Diego, CA			
Job Expert A	E-7	"A" School	7	12	19
Job Expert B	E-9	COMNAVSURFPAC	20	5	25
Job Expert C	E-9	"A" School	10	9	19
		Team 2. Alexandria, VA			
Job Expert A	01E	NAVTELCOM	5	1	6
Job Expert B	03E	USCINCLANT	13	5	18

Each of the two workshops began with an exploration of the purpose of the study and background on the previous radioman study from which the task information was obtained. The concept of an overall proficiency score was then presented, along with the task information available, which they might use to formulate this score.

The task information used came from two steps in the previous radioman study (Lammlein & Baker, 1987). One was the administration of a task survey to a sample of several hundred FTRM and their supervisors worldwide. The task information obtained from this survey included the following summary statistics:

1. Percent supervising. The percentage of FTRM supervisors responding who reported that they supervise at least one FTRM who performs the task.

2. Percent performing. The percentage of FTRM responding who reported that they perform the task.

3. Average importance rating. The average rating of task importance for mission success, as rated by FTRM supervisors on a 1-5 scale.

4. Average frequency rating. The average rating of the frequency with which the task is performed, as rated by FTRM on a 1-5 scale.

5. Average performance errors rating. The average rating of the percentage of time that the task is performed incorrectly, as rated by FTRM supervisors on a 1-5 scale.

6. Average "complicated to perform" rating. The average rating of how complicated the task is to perform, as rated by FTRM on a 1-5 scale.

For convenience of discussion, these were grouped into three "composites": (1) percentages (1 and 2); (2) importance (3 and 4); and (3) difficulty (5 and 6). The reader should note, as pointed out to the job experts, that some of the task survey information had been provided by FTRM and some by their supervisors.

The second area of task information was content area. Thirteen radioman job experts in the previous study sorted tasks into categories based on similar content. When their results were cumulated, four task groupings emerged.

1. Preparing and Processing Messages/Establishing Communications--69 tasks.
2. Setting up Equipment--23 tasks.
3. Maintaining Equipment--21 tasks.
4. Handling Secure Materials--11 tasks.

Job experts were provided with a handout summarizing the results of the job analysis. To structure the discussion of weighting systems, job experts were told by the workshop leader to proceed as follows:

1. Decide how to compute three overall composite scores using the task survey information. For example, how should average importance and average frequency values be weighted to form an overall "importance" composite value? Job experts were given the option of not using statistics that they did not deem relevant (i.e., giving such information zero weight).

2. Rank the three composites for their relative influence in weighting the overall performance score.

3. Determine the actual weights to be assigned to each composite.

4. Determine how the four content categories are to be used. For example, should a set number of tasks be drawn from each category? In essence, this was a decision of whether task sampling should be stratified using the categories.

As the job experts conducted their discussion, the workshop leader gave them feedback on the implications of the weighting systems they formulated (e.g., differences between multiplicative vs. additive weights). The determination of the final weighting system was by consensus of all the experts in each workshop.

Results

Both expert groups grasped very well the notion of an overall performance score and how the weighting of task information should be used to formulate that score. Their decisions are discussed below.

San Diego Workshop. This group determined that the three composite scores should be formed as follows:

1. Percentage = percentage performing only.
2. Importance = Average importance only.
3. Difficulty = (2 X average "complicated to perform" rating) + (1 X average performance errors rating).

These three composites should then be weighted as follows to form an overall weight for each task: (2 X Percentage) + (4 X Importance) + (1 X Difficulty). This group did not believe the content categories should be used at all in sampling tasks.

Alexandria Workshop. The three composite scores from this group were:

1. Percentage = (.45 X percentage supervising) + (.55 X percentage performing).

2. Importance = (.70 X average importance rating) + (.30 X average frequency rating).

3. Difficulty = (.90 X average performance errors rating) + (.10 X average "complicated to perform" rating).

The group then computed the overall weight for each task as follows: (.50 X Percentage) + (.25 X Importance) + (.25 Difficulty). This group also did not believe the content categories should be used in sampling tasks.

Discussion

The two groups formulated very different weighting systems. Whether this reflects a sampling effect, a workshop leader effect (each workshop had a different leader), or some other effect is difficult to discern. It may suggest that larger groups of experts should be used in the future.

Both groups did agree, however, on the fact that the content categories should not bear on task selection. Thus, both groups agreed that task sampling should be based on the entire group of 124 tasks and not stratified.

Drawing of Task Samples

The probabilistic sampling model calls for the drawing of 150 samples of 15 tasks each using the weighting system chosen by job experts. Two separate sets of 150 samples were drawn in this study, with one corresponding to each of the two weighting systems.

Method

First, a probabilistic distribution for the 124 tasks was created by each of the two weighting systems. This was accomplished by (1) standardizing the different task information measurement scales to ensure that the weightings would not be confounded by scale differences, (2) computing for each task under each weighting system the result of applying the task information as weighted by the job experts, and (3) rescaling these results with the constraint that the probability values for all 124 tasks must sum to 1.0 for each weighting system.

Using the two obtained probability distributions, 150 random samples of 15 tasks were drawn for each system using a computerized sampling procedure. In addition, 150 new task samples were drawn using each of the uniform, importance, and frequency weighting subsystems so as to provide a comparison for the job expert systems. Sampling within each task sample was without replacement of tasks previously drawn whereas sampling between task samples was with replacement.

Results

Table 5 provides summary statistics for the probability weights for the two job expert weighting systems. This table indicates that the most differentiated weighting was the job expert System 2, with job expert System 1 being more differentiated than the importance weighting but less than the frequency weighting. The weights of the two job expert systems are highly related; the correlation between them across tasks is .90.

Table 5

Summary of Probability Distributions for Uniform, Importance, Frequency, and Job Expert Weighting Systems

	Uniform	Importance	Frequency	Job Expert System 1	Job Expert System 2
Mean	.00806	.00806	.00806	.00806	.00806
SD	.00000	.00132	.00209	.00148	.00229
Maximum	.00806	.01099	.01216	.01110	.01294
Minimum	.00806	.00507	.00417	.00489	.00425
Ratio[a]	1.00	2.17	2.92	2.27	3.04

[a]Ratio of maximum to minimum probability weight.

Overlap analyses were performed on the task samples[2] drawn by the five weighting systems. These were computed as Step 1 of this study; the proportion of overlap between each task sample drawn by a given weighting and (1) the remaining 149 task samples drawn using the same weighting, and (2) the 150 task samples drawn by each of the four other weightings. Also as before, the mean importance and frequency values were computed for each task sample.

Table 6 summarizes the results. The values in this table are again the means and standard deviations of overlap proportions for task samples chosen by weighting combinations. Examination of the diagonals of this table reveals that the two job expert weighting systems resulted in task samples only trivially more homogeneous than those chosen using the uniform rating. A typical pair of task samples drawn from the uniform weighting shared 1.82 tasks, while the typical pairs from the job expert Systems 1 and 2 shared 1.85 and 1.95 tasks, respectively.

Table 6

Summary of Overlap Analyses for Different Weighting Systems

	Uniform		Importance Only		Frequency Only		Job Expert System 1		Job Expert System 2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Uniform	.121	.078								
Importance	.121	.079	.124	.080						
Frequency	.121	.079	.119	.079	.129	.081				
System 1	.122	.080	.123	.080	.122	.080	.123	.080		
System 2	.121	.080	.122	.080	.127	.081	.125	.080	.130	.082

Note. Values in this table are based on proportions of overlap between task sets.

A similar result is evident from Table 7. This table depicts the composition of the task samples drawn from each weighting in terms of importance and frequency means (from the radioman job analysis). There is a tendency for the job expert systems to select more important tasks; however, the effect is miniscule at best.

[2]Lists of the task samples are available from NAVPERSRANDCEN.

Table 7

Summary Statistics for Importance and Frequency for Task Sets Chosen by Weighting Systems

	Importance		Frequency	
	Mean	SD	Mean	SD
Uniform	3.42	.137	3.01	.197
Importance only	3.49	.123	2.97	.191
Frequency only	3.41	.144	3.19	.191
Job Expert System 1	3.47	.143	3.05	.183
Job Expert System 2	3.45	.147	3.13	.165

Discussion

One of the most interesting points to note from these results is the similarity of rank order of task selection probabilities for the two job expert weighting systems. Despite rather different weighting combinations, the probabilities from the two systems correlated .90. This result is at least partially attributable to the fact that the values for different scales from the task survey are correlated.

Just as was the case for Step 1, the weighting systems---this time including two chosen by job experts--held little constraint over the composition of the task samples. Thus, random sampling again played a considerable role in determining the final task sample.

Expert Screening of Task Samples

The probabilistic sampling model includes a screening step by job experts in which they may delete at most 10 percent of the task samples unacceptable to them for purposes of performance testing. In this study, each of the two teams of job experts reviewed the task samples drawn by its weighting system and deleted those that should not be considered further. This was accomplished first by individual review and then convening of consensus workshops. Due to the experimental nature of this study, of interest were not only the task samples deleted but also the factors the job experts considered in making their choices.

Method

With one exception, the job experts participating in this step were the same as those participating in Step 2. The only change was a substitution in the San Diego group, where Job Expert A in Table 4 was replaced due to his having left the Navy. His replacement was an E-8 Senior Chief assigned to the "A" School who had been a radioman or radioman supervisor for approximately 19 years at the time of the workshop. Of this time, 7 years had been spent on board ship and 12 had been spent at shore stations.

Job experts in each of the two groups were sent materials in advance of the workshops. These included a booklet containing the task samples and instructions for the

review task. These materials allowed each expert to review the task samples individually in advance so that a group consensus on deletions could be more easily reached in the subsequent workshop.

The instructions explained the purpose of the task--review of the task samples and deletion of at most 15 (10% of 150) that may be unacceptable. They included a discussion of reasons why a task sample might not be appropriate for performance testing (e.g., being poorly representative of the overall performance of FTRM).

To make the deletions, the instructions first called for each expert to review each task sample and rate it on a simple 3-point scale as follows:

0 = Not Acceptable
1 = Barely Acceptable
2 = Fully Acceptable

The experts were also asked to record the reason for rating a task sample as "not acceptable." Following this initial review, they were to count the task samples rated as "not acceptable." If this number was greater than 15, a second review was required to select the 15 worst task samples recommended for deletion. They were then asked to send the selections to research staff for cumulation prior to the workshop they were to attend.

In each of the two workshops, all task samples selected for deletion by at least one participant were indicated to the group. They were then instructed to reach a consensus on at most 15 task samples for deletion. Group leaders provided minimal input to this discussion, instead noting the factors used by the job experts in making their decisions.

Results

Job experts in the San Diego workshop agreed that a five-point scale for rating the samples would have been more useful than the three-point scale employed here because many samples fell between "fully acceptable" and "barely acceptable."

The "first cut" of deletions--where the number was unconstrained--will be dealt with first. Experts varied considerably in the numbers they found unacceptable. For the San Diego experts, these numbers were 27, 11, and 2. The two Alexandria experts found 59 and 23 samples unacceptable in the first cut. In terms of agreement on those selected, the San Diego participant selecting two for deletion picked one sample agreed on by the other two participants and one which was not agreed on by either. The participant initially selecting 11 samples for deletion agreed completely on those samples with the one selecting 27; that is, the 11 deleted by the first were in all cases deleted by the other. For the Alexandria participants, 12 of the 23 samples chosen by one participant were in agreement with the other.

When the San Diego expert selecting 27 task samples paired them down to 15, six of the 15 were among those selected by the participant initially selecting 11. None of the third participant's two deleted samples were included in the final selections of the first participant. For the two Alexandria experts, 12 of the 15 samples chosen in the second step were in common.

Following are the major factors that the job experts cited in making decisions about the appropriateness of task samples. These are based on the discussion in the consensus workshops. The same primary factors were used by both groups.

Over/Undersampling Job Content Areas. This seemed to be the most compelling factor to the job experts. If different areas of the job--such as preventive maintenance, equipment set-up, handling secure materials, etc.--did not receive proper emphasis in the numbers of tasks selected from each, the task sample was generally recommended for deletion. For example, some task samples were rejected because there were either not enough preventive maintenance tasks or there were too many. This was surprising in light of the fact that the job experts had earlier chosen not to include the job content areas in sampling. In addition, the content areas that the job experts seemed to use in their implicit categorization corresponded very closely to those that had earlier been identified from the card sort (Lammlein and Baker, 1987).

Inclusion of Single Tasks Not Deemed Appropriate for Performance Testing. The job experts held the view that with only fifteen tasks available for performance assessment, each task had to make some significant positive contribution to the inference of job performance. Thus, in some cases the inclusion of single tasks deemed entirely inappropriate for performance testing resulted in the entire task sample being rejected. This occurred for a number of reasons, including (1) the task is only rarely performed or is relatively trivial to mission success, (2) the task may be performed on antiquated equipment, (3) performing the task presents safety hazards, (4) the task is often performed by those in another specialty or at a higher rank, (5) the task is only relevant to certain commands, and (6) few perform poorly and thus it would contribute little in the way of assessment value to a performance test. Conversely, it was also evident that if certain tasks were included the task sample was viewed as being a stronger sample simply by virtue of including that task.

Redundant Combinations of Tasks. When some tasks occurred together, they were viewed as somewhat redundant. Thus, it was not worth using one of the 15 allowed tasks to test something similar to that tested by another task.

Using the above considerations, the job experts in San Diego agreed on the following 15 task samples for deletion from further consideration as a performance test: 32, 36, 42, 49, 62, 63, 64, 71, 72, 100, 108, 127, 128, 131, 132. The Alexandria experts deleted the following 16 (one over the number originally called for) samples: 37, 39, 53, 55, 78, 82, 87, 93, 96, 104, 116, 119 125, 145 and 147 (the reader should recall that the groups reviewed different task samples; thus, Number 59 for the San Diego group was not the same as Number 59 for the Alexandria group, and so forth). Both groups noted that they would have deleted more task samples had they been allowed to do so.

Discussion

It was clear that the job experts could successfully perform the review and deletion task. Their discussion indicates that they incorporated very relevant considerations in making their decisions.

The job experts placed great emphasis on ensuring that tasks were sampled from different job content areas. In addition, it was apparent that they had some ideal set of percentages in mind--however implicit--for what would constitute an appropriate task sample for performance testing purposes. For example, they may believe that X percent of the tasks should be preventive maintenance, Y percent security-related, and so on.

This step revealed that job experts use similar information in a similar way to that incorporated by the utility approach to task selection mentioned earlier. That is, they found the selection of certain tasks unacceptable because those tasks are infrequently performed or are relatively unimportant for mission success.

Selection of the Final Task Sample

Of the remaining task samples, one is to be chosen for purposes of performance testing. According to the probabilistic sampling model, this selection is to be made randomly. In this study, we allowed job experts a chance to comment on the selected samples.

Method

The final task sample was chosen in each of the two workshops discussed in the previous step. Of the samples remaining after the job experts had made their deletions, one was chosen using a randomly generated number. Experts then commented on it.

Results

In the San Diego workshop, task sample 92 was chosen. The experts believed that overall this sample was one of the poorer ones and would have been deleted from further consideration had they been allowed to delete more than 15 (one had, in fact deleted it on the first pass).

1. Task 17 ("Assemble/Disassemble antennas") is not a good task for performance testing.

2. Tasks 22 and 23 ("Check frequencies for usability" and "Use chirp sounder to determine frequencies", respectively) are a bit redundant.

3. Task 70 "Type messages using ACP-127 (NATO) format) is applicable only to the Atlantic fleet.

4. Task 98 ("Receive top secret materials (excluding CMS)") is ambiguous with regard to exactly what would be tested.

5. Task 21 ("Perform late starts on crypto equipment") is being phased out.

6. Equipment related tasks are overemphasized at the expense of more common message processing tasks. Those message processing tasks that are included are weak for purposes of performance testing; that is, they are not good indicators of overall performance.

7. Task 124 ("Perform preventive maintenance on telegraph telephone signal converters (using MRCs)"), the only preventive maintenance task, is most often performed by electronics technicians, not radiomen.

In the Alexandria workshop, task sample 106 was chosen as the final sample. The group considered this sample as a whole "barely acceptable". Some specific comments on it were:

1. There exists only one message preparation task; Task 37 ("Draft broadcast screen requests as appropriate."

2. Task 1 ("Set up LF transmitters") is rarely performed.

3. The sample as a whole overemphasizes security-related tasks (there are six such tasks).

4. Task 10 ("Set up the SSR-1 satellite receiver") is too easy for performance testing.

5. Task 31 ("Operate perforator") is too specific for performance testing; it is done and should be tested in conjunction with other tasks.

Discussion

Both task samples chosen in this study were viewed quite unfavorably by the job experts. In each group the experts had marked reservation about using the chosen task sample to accurately infer the performance of FTRM.

As noted earlier, the probabilistic sampling model as proposed leaves a great deal to chance in the selection of a sample of tasks for performance testing. All the task samples remaining after the experts delete 10 percent have an equal probability of selection. If this pool of task samples includes some that the experts would have liked to delete then an unacceptable sample may be chosen.

CONCLUSIONS AND RECOMMENDATIONS

This study was a productive one in that a great deal was learned about the probabilistic model for task selection. Overall conclusions and some methodological considerations are discussed below.

The Probabilistic Sampling Model is Workable

The probabilistic sampling model--including identification of a weighting system by job experts, conversion of that system to a probability distribution, drawing of 150 task samples, review of those samples and deletion of at most 10 percent that are unacceptable, and selection of a final sample at random for performance testing--is workable. There were no problems experienced with job experts being unable to carry out the research steps. However, they found the review of the task samples rather tedious and time-consuming.

The Influence of Task Information vs. Random Selection

This study demonstrated that the selection of tasks by the probabilistic sampling model was considerably influenced by the random "luck of the draw." The weighting systems chosen by job experts placed little constraint on the task samples on the basis of task information such as importance, frequency, etc.

Is this a problem? On the one hand it could be argued that the random component of task selection helps ensure the resulting sample to be as free as possible from systematic bias. On the other hand, the job experts' reactions to the samples chosen in this study indicated that a relatively large number of the samples were quite unacceptable for performance testing. This latter observation is especially critical, and it suggests that some means of more tightly controlling the outcome of the sampling would be desirable.

Some possible solutions to this problem include:

1. Stratified sampling of tasks. The job experts judged the adequacy of task samples primarily by whether each sample contained appropriate numbers of tasks from different job content areas. It is recommended that when the probabilistic sampling model is used in the future, job experts should define content areas and establish the numbers of tasks to be chosen from each a priori. In this way the drawing of task samples can be stratified.

2. Allowing deletion of more than 10 percent of the task samples if job experts see fit. It is recommended that all task samples seen as inappropriate by job experts be deleted from consideration for final selection, even if the number exceeds 10 percent of the samples.

3. Prescreening of tasks. As noted, there were some tasks that the job experts simply considered inappropriate for performance testing. It would seem useful to delete these tasks before drawing task samples provided the job experts can provide relevant justification for doing so. In addition, there may be some tasks that job experts see as especially strong for performance testing. These may be given priority in sampling.

4. Deletion of tasks from task samples. Another solution bearing mention, but not a recommendation from this study, would be to draw task samples slightly larger than needed for performance testing. Then, job experts could delete individual tasks unacceptable to them.

An important common theme of these recommendations is that they all involve lessening the influence of the random component and increasing the effect of job information and direct expert judgment on the tasks chosen for performance testing. In effect, they move the probabilistic sampling model closer in concept to the utility model.

This study prompts recommendations to decrease the role of the random component. Therefore, it is appropriate to ask: Just what is the inclusion of a random drawing of tasks accomplishing?

Conceptual Underpinnings Revisited

A core assumption of the utility model is: Performance testing is best served by including in the testing sample those aspects of the job that have the most significant consequences. The notion of consequences has generally been operationalized by including those parts of the job that are most frequently performed, on which the most time is spent, which are most important for the success of the organization, etc. What is essentially being selected with the utility model are those components of the job that have relatively high "payoff" to the organization. By assessing performance on these job components, the organization will realize the greatest utility from making decisions based on the performance testing or on other measures validated to predict these aspects of performance. Thus, the utility model of task selection results in a performance score with a straightforward conceptual meaning.

Perhaps this will also result in performance scores that are systematically biased. For example, if the parts of the job with more significant consequences are also those that are more difficult to perform, performance scores based on the sample will be lower than would be the case if one could assess performance on all aspects of the job and average the scores in some way. Thus, for comparing the performance of members of an

organization against some external standard, a task sample chosen by the utility model may be misleading. This is the case even though such a sample would effectively make relative comparisons between significant incumbents or incumbent groups with respect to worth to the organization based on task performance.

This lack of representativeness in the utility model seems to be the primary motivation for the probabilistic sampling model. This model is clearly drawn from tenets of statistical sampling theory, which emphasizes the accurate inference of population parameters from sample statistics. A key component of this inferential process is random sampling.

The probabilistic sampling model seems to have as its implicit assumption that the score of primary interest is that which would be obtained if performance were assessed on every job task that could be obtained if performance were assessed on every job task that could be required of anyone holding the job title. Of course, obtaining such a score is infeasible for most jobs, and the necessity for sampling--with the attendant problem of sampling error--raises its head. By sampling tasks randomly one attempts to minimize systematic error and thus produce an unbiased estimate of the performance score for the entire domain of tasks.

The primary conceptual problem with the probabilistic sampling model is, simply, that it works to preserve a population score that is an invalid representation of "job performance" for most jobs. It asserts that "job performance" as we wish to measure and understand it is an incumbent's performance on <u>all</u> job tasks that could conceivably be required of anyone holding that job title--thus the emphasis on not completely ruling out <u>any</u> task for selection. For most job titles, this domain of tasks represents a hodgepodge of <u>many</u> jobs all grouped under that title. Put another way, groupings of tasks for a job title tend to be so heterogeneous that they do not represent the job or the responsibilities of any one person holding that title.

It was very clear in this study that the 124 tasks do not represent the job any single FTRM in the Navy. It is the nature of the FTRM job that tasks differ across duty assignments such that some FTRM will never be required to perform some tasks. Other tasks are relatively trivial and are performed so infrequently that they are not even formally or informally trained to any FTRM except a few. Some tasks are performed by those in other specialties in some commands (e.g., electronics technicians) while FTRM perform them in others.

Given the diversity of this "domain," it is wise to sample from it so as to sample from it so as to preserve a "domain score?" Put another way, would a performance score based on a random sample from such a heterogeneous domain have any useful meaning?

Our answer to this question is negative. Job experts clearly believe that some tasks are simply not appropriate for assessing the job performance of FTRM. Random sampling only kept in tasks that they would have rejected and resulted in some task samples being viewed as quite poor for performance testing. Indeed, both "final" samples drawn in this study were viewed thus! Perhaps one of the most compelling conclusions drawn from this study is that the job experts appeared to use considerations similar to those of the utility model in evaluating the randomly drawn task samples. This demonstrates that they may not view the total set of FTRM job tasks as comprising any coherent domain in a statistical sense.

Scientists may not be the most important audience for which performance tests are constructed. A performance test must be acceptable to operational users, and this acceptance is in large part a function of face validity. If users do not believe that the performance test reflects the important parts of the job they will not use it to make decisions. It is worth considering as well that if those who are to take the performance test do not consider it a valid reflection of their job, their motivation to do well on the test may be adversely affected. Assessing job performance with task samples chosen by the probabilistic sampling model may place upon the researcher a considerable burden of proof of credibility.

Another area of acceptance should be considered as well. If the probabilistic sampling model is used to select tasks for performance testing for a diverse job, the resulting test could well include tasks that a given incumbent may never perform or be expected to perform. It would be questionable according to legal and professional guidelines to use such a measure to make personnel decisions or to serve as a criterion for validation of other measures used to make decisions (cf. Society for Industrial and Organizational Psychology, 1987; Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978).

The utility model better recognizes the complexity inherent in most jobs. It does not attempt to specify a domain a priori but instead begins with the assumption that the organization will benefit by assessing performance on those tasks that are more salient (e.g., more important, more frequently performed, etc.) to the job or collection of jobs grouped under the same title. When such tasks are selected and combined to form a test, we can assert that the scores will relate to some continuum of worth to the organization. The performance score thus has meaning, and we assert that in fact it has more meaning than one obtained using the probabilistic sampling model.

It is useful to now return to the recommendations made earlier in this chapter for future implementation of the probabilistic sampling model. Stratification, prescreening of tasks, allowing deletion of individual tasks in task samples, etc. are all attempts to either structure a very heterogeneous collection of job tasks or to make up for the adverse effects of random sampling from such a heterogeneous collection.

All these recommendations really do is attempt to include some utility model considerations to balance the possibly adverse effects of random sampling of such a heterogeneous domain (effects that are not acceptable to job experts). They in essence try to make up for the insufficient impact of differentiated weighting with task information.

There is another difficulty with the probabilistic sampling model. Inferential statistics tells us that drawing random samples will minimize the risk of drawing incorrect inferences about the population. It also tells us, in the form of expanding standard errors associated with smaller samples, that single small samples can lead us to most misleading conclusions about population parameters--despite the use of random sampling. Put another way, while we do indeed minimize risk of inaccurate inference with random sampling, for small samples this risk is still considerable.

This is a very problematic issue for performance testing, because administrative constraints seldom allow testing of more than a small number of job tasks. Assuming for the moment that it makes sense to speak of a population of tasks, samples of the sizes typical for performance testing situations may contain very little information about the population scores (i.e., the task samples may be quite unrepresentative).

In summary these difficulties plague the probabilistic sampling model:

1. The primary assumption of the model--that all tasks performed by those holding the job title represent a coherent domain from which it is useful to conceive of a domain score--is flawed given the diverse nature of all but the most simple jobs. It is quite difficult to draw a valid analogy between a population of experimental subjects and a population of tasks for most jobs.

2. Given that task samples must generally be small due to administrative constraints, these samples may be subject to considerable error in representing the entire "domain" of tasks if drawn randomly.

These two factors considerably cloud the interpretation of a performance score derived from a sample of tasks selected with the probabilistic sampling model. It would seem that the utility model produces more meaningful scores because through purposeful task selection there is a defensible relationship to a very useful construct--worth to the organization. In selecting the probabilistic sampling model over the utility model we would seem to trade an albeit biased but meaningful and user-accepted score for one that, for most jobs, most probably contains considerable error, is less acceptable to job experts, and is less defensible on all fronts except when evaluated against a basic tenet of inferential statistics, which is of dubious relevance anyway given the nature of most job domains.

In the probabilistic sampling model for selecting tasks for performance testing, we see a clear analogy to the domain sampling model for constructing psychological tests. Ghiselli, Campbell, and Zedeck (1981) noted that proponents of the domain sampling model typically begin with some concept of a domain as a collection of behaviors of interest (cf. Thorndike, 1971). Then, a test is conceived of as some carefully drawn sample from that domain.

Ghiselli et al. (1981) offered a different and probably more practical view of domain sampling. Rather than conceiving of some population of behaviors and attempting to sample from it, they instead advocated beginning with some idea of what one wishes to measure, selecting the best items possible to measure it with, and then defining the domain according to the operations selected:

> The best definition we can get of the domain or variable is given by the specific sample of items or situations that we finally select, by whatever means we do it. This is, of course, an operational definition of our trait. Consequently, it makes more sense to think of the universe (that is, the domain) not as a collection of actual items or situations, but as a hypothetical universe or population of behavior that has the same characteristics as those in our actual sample or test.
>
> Please recognize that what we have done is reversed the conventional relationship between population of some set of entities and then drawing a sample randomly from it, we have said that we have a sample in hand that implies a population (that is, a universe or domain) having the same characteristics as the sample. That is, there is a hypothetical universe of content from which our sample could have been randomly drawn. (p. 124)

This suggests that Ghiselli et al. (1981) shared some of our doubts about the usefulness of conceptualizing or constructing tests as random or stratified random samples from populations of items. It is interesting that the process they argued to be more realistic quite closely resembles the utility model for performance testing.

It is not the intention of this final section to quash further research into the probabilistic sampling model. In particular, it is hoped that future research will be conducted to try some of the methodological recommendations made earlier so as to determine the effect on task samples. It must also be noted that not all jobs are as heterogeneous as the FTRM job studied here. Some jobs no doubt have a smaller sphere of tasks; involve fewer inter-task differences in importance, frequency, etc.; present more inter-incumbent homogenity in tasks performed; train more tasks or expert competent performance on the entire set; etc. Such a job would be more appropriate for application of the probabilistic sampling model. In other words, where tasks can be assumed to be more interchangeable, the probabilistic sampling model may be more justified.

Finally, we wish to point out that the cause of science has been well served by the proposition of the probabilistic sampling model. When a new theory or method questions the conventional wisdom, assumptions heretofore implicit become explicit and open to evaluation. The probabilistic sampling model has prompted a careful look at what is being measured when performance tests are assembled using various assumptions. This is an important advance in our continuing understanding of the nature of job performance and how best to measure it.

REFERENCES

Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council. (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project, A. K. Wigsor & B. F. Green, Jr. (Eds.). Washington, DC: National Academy Press.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38294-38309.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco: W. H. Freeman.

Lammlein, S. E (1986). Proposal and evaluation of a model for job knowledge testing (Unpublished doctoral dissertation). Minneapolis: University of Minnesota.

Lammlein, S. E., & Baker, H. G. (1987). Developing performance measures for the Navy radioman (RM): Selecting critical tasks (NPRDC TN 87-13). San Diego: Navy Personnel Research and Development Center.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the validation and use of personnel selection procedures (Third Edition). College Park, MD: Author.

Thorndike, R. L. (1971). Educational measurement for the seventies. In R. L. Thorndike (ed.), Educational measurement. Washington, DC: American Council on Education.